Peertechz


ENGINEERING GROUP

TRENDS IN
Computer Science and Information Technology
OPEN ACCESS JOURNAL

ISSN: 2641-3086 DOI: https://dx.doi.org/10.17352/tcsit

Mini Review

# Visual experience recognition using adaptive support vector machine

**SP Santhoshkumar[1]*, M Praveen Kumar[1] and H Lilly Beaulah[2]**

[1]Assistant Professor, Department of IT, Rathinam Technical Campus, Coimbatore, India

[2]Professor, Department of CSE, Mahendra College of Engineering, Salem, India

**Received:** 02 September, 2021
**Accepted:** 02 December, 2021
**Published:** 03 December, 2021

***Corresponding author:** SP Santhoshkumar, Assistant Professor, Department of IT, Rathinam Technical Campus, Coimbatore, India, Tel: +91 99945 25372; E-mail: spsanthoshkumar16@gmail.com

**ORCID:** https://orcid.org/0000-0001-8531-759X

**Keywords:** Adaptive SVM; Classifiers; Kernel learning; Pyramid matching; Support vector machine



https://www.peertechzpublications.com

Check for updates

## Abstract

Video has more information than the isolated images. Processing, analyzing and understanding of contents present in videos are becoming very important. Consumer videos are generally captured by amateurs using handheld cameras of events and it contains considerable camera motion, occlusion, cluttered background, and large intraclass variations within the same type of events, making their visual cues highly variable and less discriminant. So visual event recognition is an extremely challenging task in computer vision. A visual event recognition framework for consumer videos is framed by leveraging a large amount of loosely labeled web videos. The videos are divided into training and testing sets manually. A simple method called the Aligned Space-Time Pyramid Matching method was proposed to effectively measure the distances between two video clips from different domains. Each video is divided into space-time volumes over multiple levels. A new transfer learning method is referred to as Adaptive Multiple Kernel Learning fuse the information from multiple pyramid levels, features, and copes with the considerable variation in feature distributions between videos from two domains web video domain and consumer video domain.With the help of MATLAB Simulink videos are divided and compared with web domain videos. The inputs are taken from the Kodak data set and the results are given in the form of MATLAB simulation.

## Introduction

In the past few years, computer vision researchers have witnessed a surge of interest in human action analysis through videos. With the rapid adoption of digital cameras and mobile phone cameras, visual event recognition in personal videos produced by consumers has become an important research topic due to its usefulness in automatic video retrieval and indexing. Event recognition from visual cues is a challenging task because of complex motion, cluttered backgrounds, occlusions, as well as geometric and photometric variances of objects. Previous work on video event recognition can be roughly classified as either activity recognition or abnormal event recognition. First, a large corpus of training data is collected, in the concept, labels are generally obtained through expensive human annotation. Next, robust classifiers also called models or concept detectors are learned from the training data. Finally, the classifiers are used to detect the presence of the concepts in any test data. Sufficient and strong labeled training samples are provided, these event recognition methods have achieved promising results. However, it is well-known that the learned classifiers from a limited number of labeled training samples are usually not robust and do not generalize well. This project proposes a new event recognition framework for consumer videos by leveraging a large number of loosely labeled YouTube videos. A large amount of loosely labeled YouTube can be readily obtained by using keywords-based search. YouTube videos are downsampled and compressed by the web server, so the quality of YouTube videos is generally lower than consumer videos. YouTube videos may have been selected and edited to attract attention, while consumer videos are in their natural captured state. Figure 1 shows four frames from two events picnic and sports as examples to illustrate the considerable appearance differences between consumer videos and YouTube videos. Therefore, the feature distributions of samples from the two domains web video domain and consumer video domain may



**Citation:** Santhoshkumar SP, Kumar MP, Beaulah HL (2021) Visual experience recognition using adaptive support vector machine. Trends Comput Sci Inf Technol 6(3): 072-076. DOI: https://dx.doi.org/10.17352/tcsit.000043

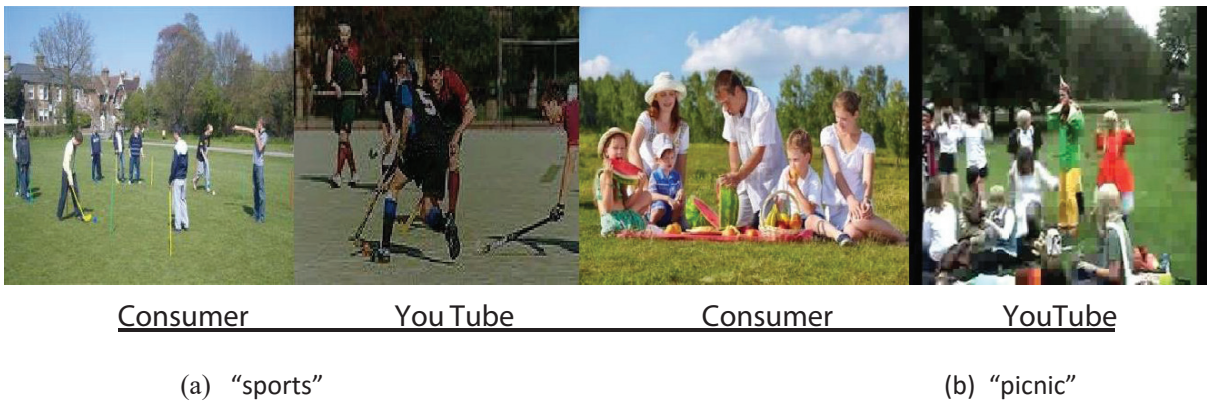


**Figure 1:** Four samples frames from consumer videos and YouTube videos. The examples from the two events illustrate the appearance differences between consumer videos and YouTube videos.

change considerably in terms of the statistical properties such as mean, intra-class, and interclass variance.

An event recognition framework extends the recent work on pyramid matching and presents a new matching method called Aligned Space-Time Pyramid Matching to effectively measure the distance between two video clips that may be from different domains. Divide each video into space-time volumes over multiple levels and calculate the pairwise distances between any two volumes and further integrate the information from different volumes with integer flow Earth Movers Distance to explicitly align the volumes. The Earth Mover's Distance (EMD) is a method to evaluate dissimilarity between two multi-dimensional distributions in some feature space. The EMD lifts this distance from individual features to full distributions.

A technique that uses local space-time features to classify six human actions like walk, jog, run, wave, clap, and box in challenging real-world video sequences. This technique achieves comparable performance in the presence of camera motion, scale variation, and viewpoint changes. Hinder the use of 2D local descriptors for object detection in static images also impact spatiotemporal local descriptors. Cross-domain learning method, referred to as Adaptive Multiple Kernel Learning (A-MKL), in order to cope with the considerable variation in feature distributions between videos from the web domain and consumerdomain. Each pyramid level and each type of local feature, train a set of Adaptive SVM classifiers.Based on a combined training set from two domains by using multiple base kernels of different kernel types and parameters, are further fused with equal weights to obtain an average classifier. A new objective function to learn an adapted classifier based on multiple base kernels and the learned average classifiers by minimizing both the structural risk functional and mismatch of data distributions from two domains.

## Related works

Event recognition methods can be roughly categorized into model-based methods and appearance-based techniques. Model-based approaches relied on various models including HMM, coupled HMM, and Dynamic Bayesian Network [1] to model the temporal evolution. Appearance-based approaches employed space-time features extracted from salient regions with significant local variations in both spatial and temporal dimensions [2-4].

Statistical learning methods including Support Vector Machine (SVM) [4], probabilistic Latent Semantic Analysis (pLSA) [3], and Boosting [5] were applied to the space-time features to obtain the final classification. Promising results [3,4,6,7,] have been reported on video data sets under controlled settings, such as Weizman [6] and KTH [4] data sets. Classifier adaptation can be seen as an effort to solve the fundamental problem of mismatched distributions between the training and testing data. This problem occurs in concept detection in a video corpus such as TRECVID [7], which contains data from different sources programs. In existing approaches [8-10], concept classifiers are built from and applied to data collected from all the programs without considering their difference in distribution. In this paper, a different scenario where classifiers trained from one or several.

## Programs are adapted to a different program are considered

The proposed classifier adaptation method is related to the work on drifting concept detection in the data mining community and transfer learning and incremental learning in the machine learning community. Incremental learning methods, such as incremental SVMs [6,11], continuously update a model with new examples without re-training over all the examples. The training and test distribution are identical, A-SVMs can be treated as a generic incremental method that can handle classifiers of any type. It is also more efficient than existing methods [6,11] whose training involves at least part of the previous examples support vectors.

## Pyramid matching

Spatial pyramid matching [8] and its space-time extension [12] used fixed block-to-block matching and fixed volume-to-volume matching. In contrast, this aligned pyramid matching extends the methods of Spatially Aligned Pyramid Matching (SAPM) [4] and Temporally Aligned Pyramid Matching (TAPM)



**Citation:** Santhoshkumar SP, Kumar MP, Beaulah HL (2021) Visual experience recognition using adaptive support vector machine. Trends Comput Sci Inf Technol 6(3): 072-076. DOI: https://dx.doi.org/10.17352/tcsit.000043

[13] from either spatial domain or temporal domain to the joint space-time domain, the volumes across different space and time locations may be matched.

Similar to [12], divide each video clip into $8^l$ non overlapped space-time volumes over multiple levels, l=0,...., L-1 where the volume size is set as $1/2^l$ of the original video in width, height, and temporal dimension. Following [12], extract the local space-time (ST) features including Histograms of Oriented Gradient (HoG) and Histograms of Optical Flow (HoF), are further concatenated together to form lengthy feature vectors. Sample each video clip to extract image frames and then extract static local SIFT features from them [10]. This method consists of two matching stages. In the first matching stage, calculate the pairwise distance $D_{rc}$ between each two space-time volumes $V_i(r)$ and $V_j(c)$, where r,c = 1,....., R with R being the total number of volumes in a video. The space-time features are vector-quantized into visual words and then each space-time volume is represented as a token-frequency feature. As suggested in [12], to measure the distance $D_{rc}$ using equation (1) Note that each space-time volume consists of a set of image blocks.

Token-frequency (tf) features from each image block are extracted by vector-quantizing the corresponding SIFT features into visual words. Based on the SIFT features, as suggested in [13], the pairwise distance $D_{rc}$ between two volumes $V_i(r)$ and $V_j(c)$ is calculated by using Earth Mover's Distance (EMD),

$$D_{rc} = \frac{\sum_{u=1}^{H}\sum_{v=1}^{I} \hat{f}duv}{\sum_{u=1}^{H}\sum_{v=1}^{I} \hat{f}uv} \quad (1)$$

Where H, I are the numbers of image blocks in $V_i(r)$, $V_j(c)$ respectively, $d_{uv}$ is the distance between two image blocks Euclidean distance is used in this work and $f_{uv}$ is the optimal flow that can be obtained by solving the linear programming problem as follows:

$$\hat{F}_{rc} = \underset{F_{rc}}{\arg\min} \sum_{r=1}^{R}\sum_{c=1}^{R} F_{rc} D_{rc} \quad (2)$$

$$S.t \sum_{c=1}^{R} F_{rc} = 1, \forall r \sum_{r=1}^{R} F_{rc} = 1, \forall c \quad (3)$$

In the second stage, further, integrate the information from different volumes with Integer-flow EMD to explicitly align the volumes. Try to solve a flow matrix $\hat{F}_{rc}$ containing binary elements that represent unique matches between volumes $V_i(r)$ and $V_j(c)$. As suggested in [4], such a binary solution can be conveniently computed by using the standard Simplex method for linear programming.

## Adaptive multiple kernel

**Learning:** The proposed framework consists of three contributions:

A visual event recognition framework for consumer videos with only a limited number of labeled consumer videos by leveraging a large amount of loosely labeled web videos.

Pyramid matching extended by presenting a new matching method called Aligned Space-Time Pyramid Matching (ASTPM) to effectively measure the distances between two video clips.

A cross-domain learning method, Adaptive Multiple Kernel Learning (A-MKL), is used to cope with the considerable variation in feature distributions between videos from the web video domain and consumer video domain by minimizing both the structural risk functional and mismatch of data distributions from two domains.

Web video domain is taken as the auxiliary domain $D^A$ source domain and the consumer video domain as the target domain $D^T$. $D^T = D^T U D^T$, Where $D^{Tl}$ and $D^T_u$ represent the labeled and unlabeled data in the target domain. Transfer learning domain adaptation or cross-domain learning methods have been proposed for many applications. To take advantage of all labeled patterns from both auxiliar y and target domains, in previous work proposed a Feature Replication (FR) by using augmented features for SVM training. In Adaptive SVM (ASVM) the target classifier $f^T(x)$ is adapted from an existing classifier $f^A(x)$ as an auxiliary classifier trained based on the samples from the auxiliary domain. Figure 2 illustrate event recognition for consumervideos by leveraging alargenumberof loosely labeled YouTube videos.

Divide each video into $8^l$ non-overlapped space-time volumes over multiple levels, l=0,..., L-1. where the volume size is set as $1/2^l$ of the original video in width, height, and temporal dimension. The partition for two videos $V_i$ and $V_j$ at level-1. The local Space-Time (ST) features including Histograms of Oriented Gradient (HoG) and Histograms of Optical Flow (HoF), are extracted and further concatenated together to form lengthy feature vectors. Sample each video clip to extract image frames and then extract static local SIFT features from them.

**The two matching stages are:** In the first matching stage, calculate the pairwise distance $D_{rc}$ between each two space-time volumes $V_i(r)$ and $V_j(c)$, where r,c=1,....., R with R being the total number of volumes in a video.

In the second stage, further, integrate the information from different volumes withInteger flow Earth Mover's Distance to explicitly align the volumes.Solve aflow matrix $\hat{F}_{rc}$ containing binary elements that represent unique matches between volumes $V_i(r)$ and $V_j(c)$ :

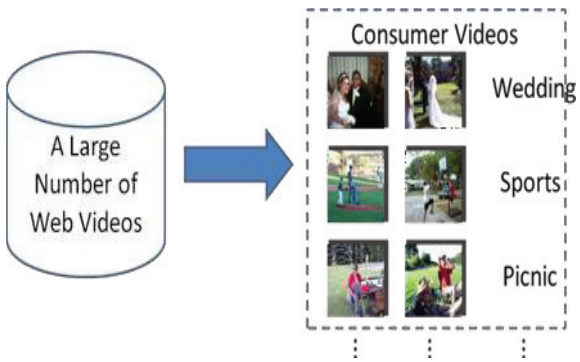


**Figure 2:** Event recognition in consumer videos by leveraging a large number of web videos.



**Citation:** Santhoshkumar SP, Kumar MP, Beaulah HL (2021) Visual experience recognition using adaptive support vector machine. Trends Comput Sci Inf Technol 6(3): 072-076. DOI: https://dx.doi.org/10.17352/tcsit.000043

$$\hat{F}_{rc} = \arg\min_{F_{rc}} \Sigma_{r=1}^{R} \Sigma_{c=1}^{R} F_{rc} D_{rc} \quad (4)$$

$$\Sigma_{u=1}^{M} f_{uv} \frac{1}{\delta}, \forall v \text{ s.t} \Sigma_{c=1}^{R} F_{\tau c} = 1, \forall r \ \Sigma_{\tau=1}^{R} F_{\tau c} = 1, \forall c \quad (5)$$

Then, the distance between two videos $V_i$ and $V_j$ canbe directly calculated by

$$D_{\Gamma}(V_i, V_j) = \frac{\Sigma_{r=1}^{R} \Sigma_{C=1}^{R} \hat{F}_{rc} D_{rc}}{\Sigma_{r=1}^{R} \Sigma_{C=1}^{R} \hat{F}_{rc}} \quad (6)$$

The matching results are obtained by using the ASTPM method. Each pair of matched volumes from two videos is highlighted in the same color. Cross-domain learning methods have been proposed for many applications [11,14,15]. To take advantage of all labeled patterns from both auxiliary and target domains, Daum´e III [14] proposed Feature Replication (FR) by using augmented features for SVM training. In Adaptive SVM (A-SVM)], the target classifier $f_T$ (x) is adapted from an existing classifier $f_A$(x) referred to as auxiliary classifier trained based on the samples from the auxiliary domain.

The target decision function is defined as While A-SVM can also employ multiple auxiliary classifiers, these auxiliary classifiers are equally fused to obtain $f_A$(x). Moreover, the target classifier $f_T$ (x) is learned based on only one kernel. Recently, Duan [15] proposed Domain Transfer SVM (DTSVM) to simultaneously reduce the mismatch in the distributions between two domains and learn a target decision function.

The learned classifiers are used prior to learning a robust adapted target classifier. Train a set of independent classifiers for each pyramid level and each type of local feature using the training data from two domains. The learned classifiers are used prior for learning a robust adapted target classifier. Further equally fuse these classifiers to obtain average classifiers $f^{SIFT}{}_{\delta}(x)$ and $f^{SIFT}{}_{\delta}(x)$ . These Classifiers are then used as prelearned classifiers $f_p(x) \parallel {}_{p=1}^{p}.\mathrm{T}$ .

The kernel function k is a linear combination of base kernels $k_m$'s, $k = \Sigma_{m=1}^{M} d_m k_m$ ,where $d_m$ is the linear combination coefficient, and the kernel function $k_m$ is induced from the nonlinear feature mapping function $\varphi_m(.)$. In A-MKL, the first objective is to reduce the mismatch in data distributions between two domains.

$$DIST_k^2(D^A, D^{\Gamma}) = \Omega(d) = h^{\delta} d \quad (7)$$

Where h = [tr($K_1$S,....,tr($K_M$S)] , and

$\varphi_m(x)\ \varphi_m(x) \in R^{NXN}$ is the mth base kernel matrix defined on the samples from both auxiliary and target domains.

The second objective of A-MKL is to minimize the structural risk functional. MKL methods utilize the training data and the test data drawn from the same domain. They come from different distributions, MKL methods may fail to learn the optimal kernel. This would degrade the classification performance in the target domain. On the contrary, A-MKL can better make use of the data from two domains to improve the classification performance.

The matching results are obtained by using the ASTPM method. Each pair of matched volumes from two videos is highlighted in the same color. The mismatch was measured by Maximum Mean Discrepancy (MMD) [16] based on the distance between the means of samples from the auxiliary domain $D_A$ and the target domain $D_T$ in the Reproducing Kernel Hilbert Space (RKHS), namely:

$$DIST_k(D^A, D^{\Gamma}) = \| \frac{1}{n_A} \Sigma_{i=1}^{n_A} \varphi(x^T{}_i) \|_H \quad (8)$$

Where $x^A$ 's and $x^T{}_i$ 's are the samples from the auxiliary and target domains, respectively. A-SVM [4,17-22] also assumes that the target classifier $f^T$ (x) is adapted from existing auxiliary classifiers. An event in consumer video is recognized using a large number of loosely labeled web videos and a limited number of labeled consumer videos. Aligned Space-Time Pyramid matching is used to find out the similarity between videos. Cross-domain learning method Adaptive Multiple Kernel Learning handles the mismatch between the data distributions of the consumer video domain and the web video domain.

## Conclusion

A new event recognition framework for consumer video is framed by leveraging a large amount of loosely labeled YouTube videos. A new pyramid matching method called ASTPM and a novel transfer learning method, A-MKL to better fuse the information from multiple pyramid levels and different types of local features and to cope with the mismatch between the feature distributions of consumer videos and web videos. A possible future research direction is to develop effective methods to select more useful videos from a large number of low-quality YouTube videos to construct the auxiliary domain.

The adaption between the web domain and consumer domain studied in this work and other examples that vision researchers have recently been working on including the adaptation of cross-category knowledge to a new category domain, knowledge transfer by mining semantic relatedness, and adaption between two domains with different feature representations. In the future, this method will be extended to A-MKL for internet vision applications.

## References

1. Hu Y, Cao L, Lv F, Yan S, Gong Y, et al. (2009) Action Detection in Complex Scenes with Spatial and Temporal Ambiguities. Proc 12th IEEE IntConf. Computer Vision 128-135. Link: https://bit.ly/3DdjGa7

2. Lazebnik S, Schmid C, Ponce J (2006) Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories. Proc IEEE Conf Computer Vision and Pattern Recognition 2169-2178. Link: https://bit.ly/3dbAvI7



Citation: Santhoshkumar SP, Kumar MP, Beaulah HL (2021) Visual experience recognition using adaptive support vector machine. Trends Comput Sci Inf Technol 6(3): 072-076. DOI: https://dx.doi.org/10.17352/tcsit.000043

3. Duan L, Tsang IW, Xu D, Maybank SJ (2009) Domain Transfer SVM for Video Concept Detection. Proc IEEE Int Conf Computer Vision and Pattern Recognition. Link: https://bit.ly/3G3JLKM

4. Duan L, Xu D, Tsang IW, Luo J (2010) Visual Event Recognition in Videos by Learning from Web Data. Proc IEEE Int Conf Computer Vision and Pattern Recognition. Link: https://bit.ly/3rBnRdE

5. Gorelick L, Blank M, Shechtman E, Irani M, Basri R (2005) Actions as Space-Time Shapes. Proc 10th IEEE Int Conf Computer Vision 29: 1395-1402. Link: https://bit.ly/3I9j2hI

6. Brand M, Oliver N, Pentland A(1997) Coupled Hidden Markov Models for Complex Action Recognition. Proc IEEE Conf Computer Vision and Pattern Recognition 994-999. Link: https://bit.ly/3dhGo6p

7. Borgwardt KM, Gretton A, Rasch MJ, Kriegel HP, Scho¨lkopf B, et al. (2006) Integrating Structured Biological Data by Kernel Maximum Mean Discrepancy. Bioinformatics 22: e49- e57. Link: https://bit.ly/3dru6ZD

8. Blitzer J, McDonald R, Pereira F(2006) Domain Adaptation with Structural Correspondence Learning. Proc Conf Empirical Methods in Natural Language 120-128. Link: https://bit.ly/3G801dC

9. Chang SF, Ellis D, Jiang W, Lee K, Yanagawa A, et al. (2007) Large-Scale Multimodal Semantic Concept Detection for Consumer Video. Proc ACM Int'l Workshop Multimedia Information Retrieval 255-264. Link: https://bit.ly/31gocYh

10. Hays J, Efros AA (2007) Scene Completion Using Millions of Photographs. ACM Trans Graphics 26. Link: https://bit.ly/31I0CtQ

11. Daume III H(2007) Frustratingly Easy Domain Adaptation. Proc Ann Meeting Assoc for Computational Linguistics 256-263. Link: https://bit.ly/3G4Cevc

12. Ke Y, Sukthankar R, Hebert M(2005) Efficient Visual Event Detection Using Volumetric Features. Proc 10th IEEE Int Conf Computer Vision 1: 166-173. Link: https://bit.ly/3G82Ifq

13. Loui AC, Luo J, Chang SF, Ellis D, Jiang W, et al. (2007) Kodak's Consumer Video Benchmark Data Set: Concept Definition and Annotation. Proc Int Workshop Multimedia Information Retrieval 245-254. Link: https://bit.ly/3EkORBS

14. Jensen PA, Bard JF (2003) Operations Research Models and Methods. John Wiley and Sons 700. Link: https://bit.ly/3rtPipO

15. Kwok JT , Tsang IW(2003) Learning with Idealized Kernels. Proc Int'l Conf Machine Learning 400-407. Link: https://bit.ly/3rt27Rt

16. Chang CC, Lin CJ (2001) LIBSVM: A Library for Support Vector Machines. Link: https://bit.ly/31kbEz6

17. Laptev I, Lindeberg T (2003) Space-Time Interest Points. Proc IEEE Int'l Conf Computer Vision 432-439. Link: https://bit.ly/3d9mHO7

18. Lanckriet GRG, Cristianini N, Bartlett P, El Ghaoui L, Jordan MI (2004) Learning the Kernel Matrix with Semidefinite Programming. J Machine Learning Research 5: 27-72. Link: https://bit.ly/3dac5yu

19. Dolla P, Rabaud V, Cottrell G, Belongie S (2005) Behavior Recognition via Sparse Spatio-Temporal Features. Proc IEEE Int Workshop Visual Surveillance and Performance Evaluation of Tracking and Surveillance. 65-72. Link: https://bit.ly/3oel3RT

20. Grauman K, Darrell T (2005) The Pyramid Match Kernel: Discriminative Classification with Sets of Image Features. Proc 10th IEEE Int'l Conf Computer Vision 1458-1465. Link: https://bit.ly/3DgGAO6

21. Laptev I, Marszałek M, Schmid C, Rozenfeld B (2008) Learning Realistic Human Actions from Movies. Proc IEEE Conf Computer Vision and Pattern Recognition 1-8. Link: https://bit.ly/3xKUhDD

22. Ikizler-Cinbis N, Cinbis RG, Sclaroff S (2009) Learning Actions from the Web. Proc 12th IEEE Int'l Conf Computer Vision 995-1002.






**Citation:** Santhoshkumar SP, Kumar MP, Beaulah HL (2021) Visual experience recognition using adaptive support vector machine. Trends Comput Sci Inf Technol 6(3): 072-076. DOI: https://dx.doi.org/10.17352/tcsit.000043